A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models

Year: 2024 | Citations: 546 | Authors: Wenqi Fan, Yujuan Ding, Liang-bo Ning, Shijie Wang, Hengyun Li

Abstract

As one of the most advanced techniques in AI, Retrieval-Augmented Generation (RAG) can offer reliable and up-to-date external knowledge, providing huge convenience for numerous tasks. Particularly in the era of AI-Generated Content (AIGC), the powerful capacity of retrieval in providing additional knowledge enables RAG to assist existing generative AI in producing high-quality outputs. Recently, Large Language Models (LLMs) have demonstrated revolutionary abilities in language understanding and generation, while still facing inherent limitations such as hallucinations and out-of-date internal knowledge. Given the powerful abilities of RAG in providing the latest and helpful auxiliary information, Retrieval-Augmented Large Language Models (RA-LLMs) have emerged to harness external and authoritative knowledge bases, rather than solely relying on the model's internal knowledge, to augment the quality of the generated content of LLMs. In this survey, we comprehensively review existing research studies in RA-LLMs, covering three primary technical perspectives: Furthermore, to deliver deeper insights, we discuss current limitations and several promising directions for future research. Updated information about this survey can be found at: https://advanced-recommender-systems.github.io/RAG-Meets-LLMs/